Auto-Encoder Transposed Permutation Importance Outlier Detector

Dr. Eren Unlu

Paris, France
datascientist.unlu@gmail.com

***Abstract*—We propose an innovative, trivial yet effective unsupervised outlier detection algorithm called *Auto-Encoder Transposed Permutation Importance Outlier Detector* (ATPI), which is based on the fusion of two machine learning concepts, auto-encoders and permutation importance. As unsupervised anomaly detection is a subjective task, where the accuracy of results can vary on the demand; we believe this kind of a novel framework has a great potential in this field.**

I. Introduction

Unsupervised anomaly detection has been one of the most extensively studied field of machine learning due to its important diverse real life applications [1]. These include fraud detection, automated identification of malfunctioning computer servers, medical diagnosis, intrusion detection and many more [2]. One particular interesting feature of this field is about the fact that there is no well defined specific accuracy method or metric as there is no supervision. The outcomes of the algorithms also depend on the perspective of the user and highly subjective. It is obvious that there is no clear definition of anomaly under an unsupervised setting, where two different anomalous instances identified by two different algorithms may be both correct under different contexts [3] [4]. For instance, the definition of anomaly for server malfunctioning shall be deviant from the fraud detection, where same algorithm may not comply the needs of both [4]. The first probable occurrence of a formal terminology for such problem in the literature is in the seminal paper of Grubbs in 1969 [5] [3].

This ambiguity of the subject make it a more attractive area for researchers as various dissimilar algorithms can be developed with varying semantic evaluation of the users [6]. As the number of features, in other words the dimension of the problem increases, the semantic evaluation of the algorithms' performance becomes more and more difficult for humans to interpret. The curse of extremely large dimensional setting also forces researchers to develop more elaborate and interpretive solutions. The general a priori assumption in unsupervised anomaly detection is that the user knows more or less the ratio of minority anomalous instances in the dataset, which is called *contamination*, where it is given as a common parameter to algorithms [7] [8].

The unsupervised anomaly detection algorithms in the literature can be grouped in to three broad categories; *proximity based*, *clustering based* and *statistical modeling based* methods [9]. Albeit most of the well known methods in the literature fall either in one of these three groups or their intersections, there also exists certain types of algorithms which can not be explained fully with this taxonomy [9]. Proximity based algorithms characterize each point with their position in the feature space with regard to their closest neighbors. By defining a proper distance metric, either the density of the data points in the vicinity or a direct distance based measure is used the score the anomaly of the point of interest in this neighborhood [10] [11]. On the other hand, clustering based algorithms aim to group data points in the feature space either directly based on the values or transformed metrics such as explaining the local connectivity of an instance [3]. In an iterative or single step fashion, the clustering algorithms encapsulates the most anomalous points in one minority class whose size is determined by the contamination ratio given by the user. Finally as the name suggests, statistical models tries to fit distributions or statistical systems to assign highest anomaly scores to a subgroup of pre-defined contamination size based on the inherent attributes of the data.

In this paper, we present an innovative unsupervised anomaly detection algorithm, where it is difficult to place categorically into this tertiary taxonomy. Our method is actually highly straight-forward and trivial but very effective to provide more intuitive alternative solutions, especially under high dimensionality. It encompasses two seminal concepts of machine learning *auto-encoders* and *permutation importance* with a simple data manipulation trick, where the dataset is transposed before being fed to algorithm as input.

II. Proposed Method

We are mainly inspired by the potency of the one of the hottest topics of machine learning, interpretability algorithms. Most particularly, permutation importance algorithm which is very straight-forward, trivial yet highly effective for global interpretability [12] [13]. The central idea of the algorithm is to measure the degree of variance of the result of the classifier or regressor for each feature independently by randomly shuffling the data points on that particular axis. Conveniently, if the result does not change significantly, it is concluded that the feature of interest has no great importance. Thus, based on

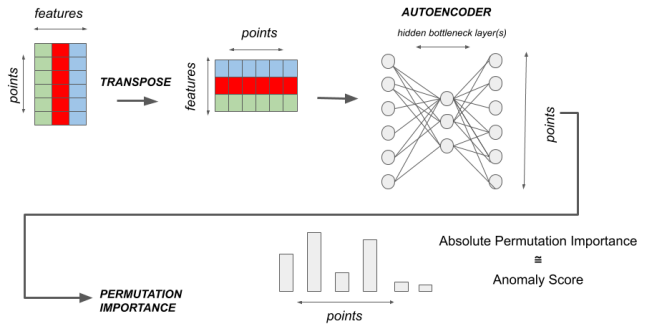

Fig. 1. Workflow of the proposed ATPI outlier detection algorithm.

this rationale each feature can be assigned relative scores of importance.

Another highly important recent topic in machine learning are the auto-encoders [14]. Albeit being a concept proposed in mid 80s by Rumelhart and Hinton in [15], their potency has been surfaced recently in parallel with the proliferation of practical deep learning architectures thanks to computational power and data adequacy. Today, they are at the core of numerous technological breakthroughs in various fields such as signal denoising, data compression, statistical modeling, information retrieval and yet anomaly detection [16] [17]. Artificial Neural Networks (ANNs) are exceptional machine learning models with their ability to learn highly complex non-linear features, inspired by biological neural structures. Current course of artificial intelligence still follows the deepening of similar architectures with advancing silicon technology and data availability. Auto-encoders are special application of ANNs for unsupervised learning where the the data of interest constitutes both input and output during training [18]. The central idea is to design an *hourglass* architecture, where at first half of the layers the number of neurons gradually decrease, finally reaching a *bottleneck layer* and in following expand back in symmetry towards the output [17]. This scheme allows the concentration of the most informative representations on the bottleneck. Therefore, for instance, the feature vector on this layer which is shorter than the actual feature dimension can be used for efficient compression or pattern recognition. In consequence, the output of these networks contain the same data points in dimension, whereas such a version excluded from outliers and noise [19] [18]. Hence, deep auto-encoders have been investigated extensively for unsupervised anomaly detection recently, providing some important improvements on the topic [20] [19].

We propose a highly trivial, yet efficient algorithm fusing these two machine learning concepts which is capable of providing insightful results for unsupervised anomaly detection problem, especially with very high dimensional dataset of relatively smaller number of samples (*wide dataframes*). The procedure is illustrated in Fig. 1. Firsly, we apply a simple trick, where the data frame is transposed. Next, an auto-encoder architecture for this transposed dataset is constructed. As it can be seen, the number of input and output neurons in this case is equal to the number of samples, not the features. Next, the network is auto-trained where number of data points of training in this case equals to the number of features. After adequate training, the outputs of this auto-encoder network assumed to contain the most informative, stripped of anomalies of this transposed dataset. Finally, a permutation importance algorithm is applied on the network, where the weights correspond to the importance of data points, not the features as the input is transposed.

As the permutation importance measures the importance of a feature by shuffling the data points randomly and checking the effect on the output, in this specific transposed case the resulting weights shall be correlated to the anomaly score of each point. Note that, as the input is transposed, the weight vector of permutation importance algorithm is equal to the number of data points, a weight correlated to the anomaly of each data point (*Fig. 1*).

III. Experimental Evaluation

We have tested a single hidden bottleneck layer architecture where the number of neurons on this layer is equal to half of the input size, i. e. number of datapoints in this specific case. 100 random shuffles are performed for permutation importance measurements. As a baseline algorithm for comparison widely known isolation forest is used [21]. 100 parallel estimators are used for isolation forest, where all features are used for training each of them. We have tested the performance of the proposed method with two well known datasets, *Wine* and *Boston* [22]. The data is scaled between 0.0 and 1.0 before being fed to algorithms for each feature. Wine dataset is composed of 178 samples of 3 classes of different wines (from 3 different regions) and their 13 numerical features. Without loss of generality, we have one hot encoded the wine type and included in the features. Thus, at the end dataset had 16 numerical features. Boston dataset is composed of 60 different houses' prices in Boston, USA and 13 numerical socio-economic, location and demographic attributes. We have included the price, thus at the end there are 14 features.

For the wine dataset we have set a contamination rate of 15%. Fig. 2, Fig. 3, Fig. 4 and Fig. 5 show the comparison of detected outlier points with two pairs of features for the proposed ATPI algorithm and isolation forest. The outlier points are marked with red. For interpretability we have also marked type of wines with three different shapes.

If we look closely to Fig. 2, we can see that in this high dimensional setting, even though both algorithms identify a handful number of common points as outliers, there are also many different points where one of them recognizes as an anomaly whilst other not. It is interesting to observe that ATPI classifies 3 neighboring points (on this specific two dimensional space, *alcalinity of ash* versus *alcohol*) which have an *alcalinity of ash* around 25.0 and *alcohol* between 13.5 and 14.5 (two of them being of class-2 and one of them belonging to class-1); isolation forest only detects one of them

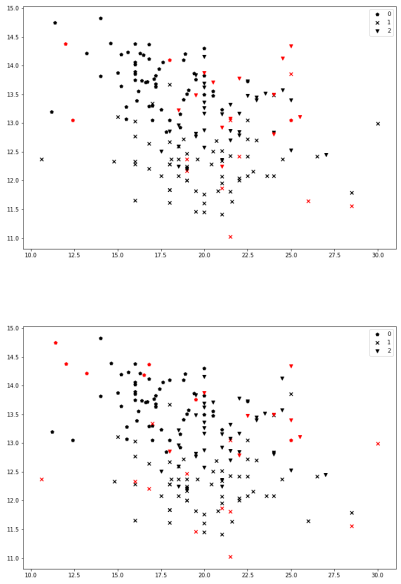

Fig. 2. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *alcalinity of ash* (x-axis) versus *alcohol* (y-axis) in *Wine* dataset

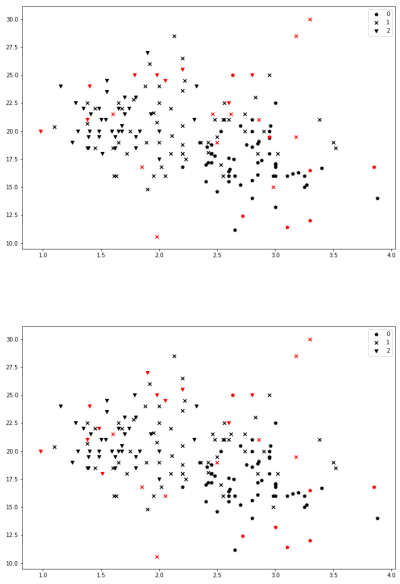

Fig. 4. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *total phenols* (x-axis) versus *alcalinity of ash* (y-axis) in *Wine* dataset

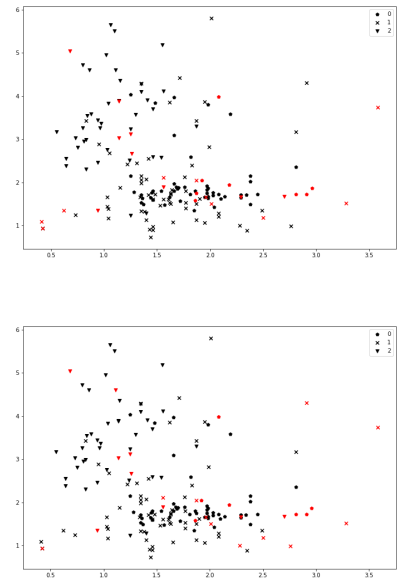

Fig. 3. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *proanthocyanins* (x-axis) versus *malic acid* (y-axis) in *Wine* dataset

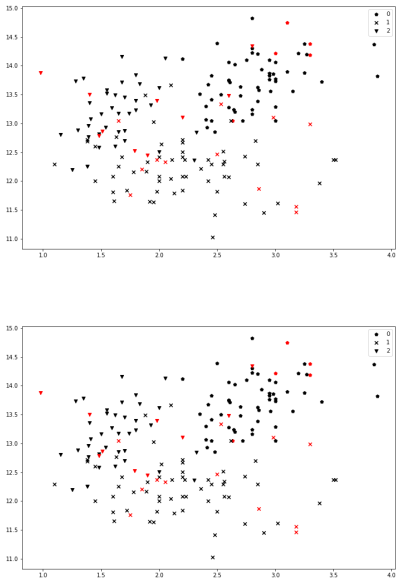

Fig. 5. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *total phenols* (x-axis) versus *alcohol* (y-axis) in *Wine* dataset

as an outlier. On the other hand, 3 neighboring points which have an *alcalinity of ash* between 11.0 and 13.5 and *alcohol* between 14.0 and 15.0 (all of them being of class-0) are all identified as outlier by isolation forest, however ATPI detected only one of them. This again demonstrates the ability of a new kind of algorithm to provide a different and valuable perspective for an unsupervised task. On Fig. 3, this time for *proanthocyanins* versus *malic acid* we observe similar outcomes. Many points are both identified as anomalies by two algorithms. However, we see that our proposed algorithm is able to classify 2 neighboring wine samples of class-1 on the bottom left of the graph as outlier, while isolation forest

can only identify one of them.

Similar observations can be made on Fig. 4 and Fig. 5. As mentioned previously, unsupervised anomaly detection, especially with large dimensionality does not permit a unified definition or measure of success, where it depends on the context and the evaluation of the human interpreter. However, it is highly encouraging to see that ATPI can identify many common anomalies with a baseline algorithm, whereas it is still able to provide different meaningful anomalous samples. This proves the ability of the proposed algorithm to give a different, novel perspective compared to conventional algorithms with a reasonable confidence.

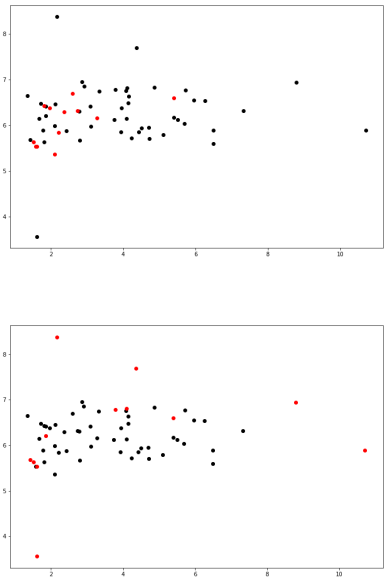

Fig. 6. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *DIS* (x-axis) versus *RM* (y-axis) in *Boston* dataset

Experiments on *Boston* dataset demonstrate similar outcomes for the *Boston* dataset, with the exception for Fig. 6 (*DIS* versus *RM*). In this specific case, all apparent outliers are missed by ATPI, whereas isolation forest managed to detect. However, note that our algorithm has been able to identify a pattern; a relatively large cluster of neighboring data points on the left handside of the graph are marked as anomaly. Also, note there still exists a significant number of common data points identified by both of the algorithms. This hints about the potent anomalous pattern recognition capability of our algorithm, with a different perspective on the issue compared to a conventional method. Especially, in Fig. 8 (*TAX* versus *PTRATIO*), we see that ATPI can detect two apparent outliers on the top of the graph, while isolation forest misses.

IV. Conclusion and Perspectives

We have developed a new kind of unsupervised outlier detection algorithm called *Auto-Encoder Transposed Permutation Importance Outlier Detector* (ATPI), which integrates

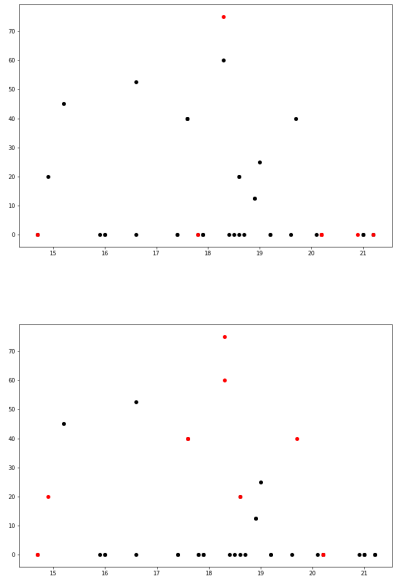

Fig. 7. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *PTRATIO* (x-axis) versus *ZN* (y-axis) in *Boston* dataset

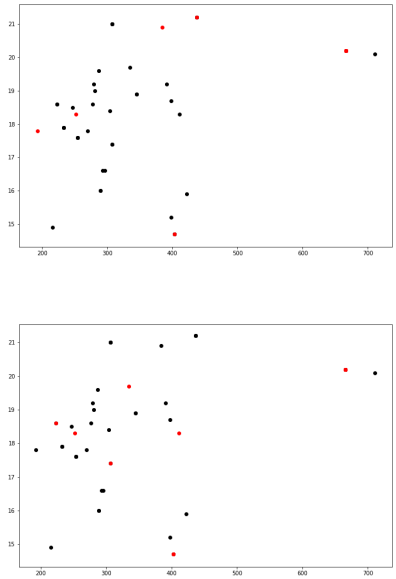

Fig. 8. The detected outlier points with proposed ATPI algorithm (top graph) and isolation forest (bottom graph) for *TAX* (x-axis) versus *PTRATIO* (y-axis) in *Boston* dataset

two potent concepts of machine learning; auto-encoders and permutation importance interpretability method. Our algorithm depends on a simple yet effective trick where a deep auto-encoder is trained by the transposed dataset. In other words, the features are treated as samples and vice versa. Thus, when a permutation importance is applied on this network,

the resulting feature importance weights can be treated as a measure of data point anomaly. Unsupervised outlier and novelty detection is a highly interesting area as there is no universal definition of accuracy and the performance depends on the context and semantic interpretation of users. The ambiguity and challenge increases as the number of features increases, where efficient visual human interpretation in high dimensional space is not possible. Therefore, introduction of new types of novel algorithms based on different mechanisms has a colossal importance.

It was demonstrated with experiments on relatively high dimensional datasets that ATPI can identify numerous common data points as outliers, which are also detected by a conventional algorithm. Whilst, it still suggests different but semantically meaningful anomalous points; which is an indicator on the capabilities of the proposed method. We believe that the introduced framework in this paper has a great potential in unsupervised novelty detection. More elaborate solutions can be developed based on this paradigm by using different kinds of auto-encoder networks and machine learning interpretation algorithms.

REFERENCES

[1] R. Ramya and G. S. Krishna, "Design and analysis of autonomous anomaly detection models."

[2] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection," in *Applications of data mining in computer security*. Springer, 2002, pp. 77–101.

[3] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[4] C. Fan, F. Xiao, Z. Li, and J. Wang, "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review," *Energy and Buildings*, vol. 159, pp. 296–308, 2018.

[5] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[6] P. Gogoi, D. Bhattacharyya, B. Borah, and J. K. Kalita, "A survey of outlier detection methods in network anomaly identification," *The Computer Journal*, vol. 54, no. 4, pp. 570–588, 2011.

[7] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 5, no. 5, pp. 363–387, 2012.

[8] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708–713, 2015.

[9] M. Goldstein and A. Dengel, "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," *KI-2012: Poster and Demo Track*, pp. 59–63, 2012.

[10] M. J. Prerau and E. Eskin, "Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm," *Undergraduate Thesis, Columbia University: December*, 2000.

[11] F. Falcão, T. Zoppi, C. B. V. Silva, A. Santos, B. Fonseca, A. Ceccarelli, and A. Bondavalli, "Quantitative comparison of unsupervised anomaly detection algorithms for intrusion detection," in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019, pp. 318–327.

[12] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[13] N. Huang, G. Lu, and D. Xu, "A permutation importance-based feature selection method for short-term electricity load forecasting using random forest," *Energies*, vol. 9, no. 10, p. 767, 2016.

[14] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[16] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[17] W.-J. Jia and Y.-D. Zhang, "Survey on theories and methods of autoencoder," *Computer Systems & Applications*, no. 5, p. 1, 2018.

[18] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[19] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 665–674.

[20] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[21] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[22] T. Hauck, *scikit-learn Cookbook*. Packt Publishing Ltd, 2014.